Bayesian Calibration of p-Values from Fisher's Exact Test

Ott, Manuela ; Held, Leonhard

Abstract: p-Values are commonly transformed to lower bounds on Bayes factors, so-called minimum Bayes factors. For the linear model, a sample-size adjusted minimum Bayes factor over the class of g-priors on the regression coefficients has recently been proposed (Held Ott, The American Statistician 70(4), 335–341, 2016). Here, we extend this methodology to a logistic regression to obtain a sample-size adjusted minimum Bayes factor for 2×2 contingency tables. We then study the relationship between this minimum Bayes factor and two-sided p-values from Fisher's exact test, as well as less conservative alternatives, with a novel parametric regression approach. It turns out that for all p-values considered, the maximal evidence against the point null hypothesis is inversely related to the sample size. The same qualitative relationship is observed for minimum Bayes factors over the more general class of symmetric prior distributions. For the p-values from Fisher's exact test, the minimum Bayes factors do on average not tend to the large-sample bound as the sample size becomes large, but for the less conservative alternatives, the large-sample behaviour is as expected.

DOI: https://doi.org/10.1111/insr.12307

Bayesian calibration of p-values from Fisher's exact test

Manuela Ott and Leonhard Held

Epidemiology, Biostatistics and Prevention Institute
University of Zurich
Hirschengraben 84, 8001 Zurich, Switzerland
E-mail: manuela.ott@uzh.ch, leonhard.held@uzh.ch

April 4, 2018

p-Values are commonly transformed to lower bounds on Bayes factors, so-called minimum Bayes factors. For the linear model, a sample-size adjusted minimum Bayes factor over the class of g-priors on the regression coefficients has recently been proposed (Held & Ott, The American Statistician 70(4), 335-341, 2016). Here we extend that methodology to a logistic regression to obtain a sample-size adjusted minimum Bayes factor for 2×2 contingency tables. We then study the relationship between this minimum Bayes factor and two-sided p-values from Fisher's exact test, as well as less conservative alternatives, with a novel parametric regression approach. It turns out that for all p-values considered, the maximal evidence against the point null hypothesis is inversely related to the sample size. The same qualitative relationship is observed for minimum Bayes factors over the more general class of symmetric prior distributions. For the p-values from Fisher's exact test, the minimum Bayes factors do on average not tend to the large-sample bound as the sample size becomes large, but for the less

conservative alternatives the large-sample behavior is as expected.

Key Words: Evidence, Fisher's exact test, g-prior, Minimum Bayes factor, Objective Bayes, p-value, Sample size, 2×2 contingency tables

1 Introduction

p-Values are indirect measures of evidence, they quantify the degree of conflict between the null hypothesis H_0 and the data (Goodman, 1992). However, if an alternative hypothesis H_1 has also been specified, the evidence for the alternative H_1 against the null hypothesis H_0 (or vice versa) is usually of primary interest. p-Values can be transformed to Bayes factors, which directly measure the evidence of the data for H_1 against H_0 (or vice versa). If the alternative H_1 is composite, it is common to derive a lower bound on the Bayes factor over a specific class of prior distributions on the parameter of interest under H_1 - a so-called minimum Bayes factor (minBF) - to provide a more objective measure of evidence and to avoid elicitation of all parameters of the prior distribution for that parameter. It is a well-known finding that for point null hypotheses, minimum Bayes factors provide less evidence against H_0 than the corresponding p-value might suggest (Berger & Sellke, 1987; Sellke et al., 2001). Ghosh et al. (2005) and Held & Ott (2018) review the literature on minBFs and transformations of p-values to minBFs.

Most Bayesian calibrations of p-values in the literature transform a given p-value to the same minBF no matter what the underlying sample size is. However, the evidence of a p-value is known to depend on the sample size (Royall, 1986; Spiegelhalter et al., 2004; Wagenmakers, 2007). Held & Ott (2016) derived a sample-size adjusted minBF over the class of g-priors for F-test p-values in the linear model. They found that the maximal evidence of these p-values is inversely related to sample size. An interesting question is whether the same is true in generalized linear models (GLMs) and, if

so, how strong the dependence on sample size is. In this paper, we extend their methodology to a logistic regression to obtain a sample-size adjusted minBF for 2×2 contingency tables. We will then relate that minBF to two-sided p-values from Fisher's exact test to quantify how their maximal evidence depends on sample size.

Let β denote the log odds ratio for an outcome of interest between two groups compared, usually estimated from the corresponding 2×2 table. In the sequel, we consider testing the point null hypothesis H_0: $\beta = 0$ against either the two-sided alternative H_1: $\beta \neq 0$ or the one-sided alternative H_1: $\beta < 0$. One-sided alternatives are of interest since practitioners often have a strong prediction about the direction of the effect if the alternative is true. A traditional rule is to apply Fisher's exact test instead of the chi-squared test if some of the expected frequencies of the 2×2 table are less than 5 (Fisher, 1941), see Andrés & Tejedor (1997) for more differentiated recommendations. However, since Fisher's exact test is conservative (D'Agostino et al., 1988; Hirji et al., 1991), the associated p-values tend to be biased in an upward direction (Barnard, 1989). We will thus also consider less conservative alternatives to these p-values such as the mid p-value.

The Bayes factor for the null hypothesis H_0 against the alternative H_1 is given by

$$\mathrm{BF}_{01} = \frac{f(\text{data} \mid H_0)}{f(\text{data} \mid H_1)},$$

where $f(\text{data} \mid H_0)$ is the likelihood of the data under H_0 and $f(\text{data} \mid H_1)$ the marginal likelihood under H_1. The denominator $f(\text{data} \mid H_1)$ depends on the prior distribution for β and we will maximize that marginal likelihood within a certain class of prior distributions (including the point distribution at H_0) to get an objective lower bound on the Bayes factor BF_{01}. This lower bound is never larger than one, so lies in the same range as the p-value, which facilitates comparisons.

This paper is structured as follows: In the next section, several large-sample minBFs

from the literature are reviewed. Some of them will play an important role when we study the asymptotic behavior of the proposed sample-size adjusted minBFs. In Section 3, one- and two-sided p-values from Fisher's exact test and alternative, less conservative significance measures (a mid p-value and a Bayesian significance measure from Liebermeister's test, see Section 3.3 for the latter) are introduced and applied to an illustrative example in Section 3.4. Sample-size adjusted minBFs for GLMs under local normal alternatives are described in Section 4, where we establish a new convergence result for the minBF (Section 4.2). The proposed sample-size adjusted minBFs for 2×2 tables are presented in Section 5. We consider minBFs over two classes of prior distributions on the log odds ratio under the two-sided alternative hypothesis: First, a class of normal priors (Section 5.1), where we apply the methodology introduced in Section 4, and then a more general class of symmetric priors (Section 5.2). The relationship between p-values from Fisher's exact test and the proposed minBFs is investigated in detail in Section 6 based on a novel parametric regression approach. The same relationship is also studied for the less conservative alternatives to these p-values. It turns out that the maximal evidence of all these p-values/significance measures against the point null hypothesis is inversely related to the sample size. The summary points in Section 6 highlight our main findings. We close with some discussion in Section 7.

2 Large-sample minimum Bayes factors

To begin, we focus on Bayes factors based on the sampling distribution of test statistics instead of the original data (Johnson, 2005, 2008). For GLMs, the likelihood ratio test statistic (or deviance) is a common choice. As is well-known, under some regularity conditions, the asymptotic distribution of the deviance is a central chi-squared distribution under the null hypothesis and a non-central chi-squared distribution under the alternative (Davidson & Lever, 1970). From these results, the so-called test-based

Bayes factor can be derived under the assumption of a generalized g-prior (Held et al., 2015), *i.e.* a (multivariate) normal prior centered around $\mathbf{0}$, on the vector of regression coefficients (see Section 4.1 for more information on generalized g-priors). Let $z = z(p) = Q_{\chi^2(d)}(1-p)$ be the deviance, where $Q_{\chi^2(d)}(.)$ denotes the quantile function of the χ^2-distribution with d degrees of freedom. The minimum test-based Bayes factor based on the deviance z is then (Johnson, 2008)

$$\text{minTBF}_d(p) = \begin{cases} \left(\frac{z}{d}\right)^{d/2} \exp\left(-\frac{z-d}{2}\right) & \text{for } z = z(p) > d \\ 1 & \text{otherwise,} \end{cases} \tag{1}$$

and has been shown to be equivalent to well-known large-sample minBFs from the literature for $d = 1$, $d = 2$ and $d \to \infty$ (Held et al., 2015). For a 2×2 contingency table with fixed margins, we have $d = 1$, so that we are primarily interested in this case. The corresponding large-sample minBF can be written as

$$\text{minTBF}_1(p) = \begin{cases} \sqrt{z}\, \exp(-z/2)\sqrt{e} & \text{for } z = z(p) > 1 \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

However, the case $d = 2$ is also of interest because it corresponds to a popular and easily applicable minBF which directly calibrates a two-sided p-value p as follows (Vovk, 1993, section 9):

$$\text{minTBF}_2(p) = \begin{cases} -e\, p \log(p) & \text{for } p < 1/e \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

A simple derivation of (3) can be found in Sellke et al. (2001) and is outlined in Held & Ott (2016, appendix B). This minBF will be termed the "$-ep\log(p)$" calibration in the following. For normally distributed data, the bound (3) has been shown to lie close to the minBF over the class of all unimodal prior distributions on the mean μ which

are symmetric around the null value (*i. e.* the value under the point null hypothesis H_0) $\mu = 0$ (Sellke et al., 2001). That class "is often argued to contain all objective and sensible priors" (Sellke et al., 2001).

For a one-sided p-value p, the case $d \to \infty$ is also relevant. The ''Edwards" (Edwards et al., 1963) bound is given by

$$\text{minTBF}_\infty(p) = \begin{cases} \exp(-t^2/2) & \text{for } p < 0.5 \\ 1 & \text{otherwise,} \end{cases} \tag{4}$$

where $t = \Phi^{-1}(1-p)$ is the one-sided z-value. This bound is obtained if we consider a normally distributed observation $y \sim \text{N}(\mu, \sigma^2)$ with known variance σ^2 and test H_0: $\mu = \mu_0$, assuming the class of all possible prior densities on μ under the alternative H_1. The lower bound is then attained for the simple alternative H_1: $\mu = \mu_1 = y$.

Another calibration, which is directly based on the two-sided p-value p and can be derived along similar lines as the "$-ep\log(p)$" calibration (3) (Held & Ott, 2018, section 2.3), is

$$\text{minBF}_\infty(p) = \begin{cases} -e\,(1-p)\log(1-p) & \text{for } p < 1 - 1/e \\ 1 & \text{otherwise.} \end{cases} \tag{5}$$

This bound is obtained by replacing p in (3) by $q = 1 - p$ and will thus be called the "$-eq\log(q)$" calibration in the sequel. Note that calibration (5) is a much lower bound than the minimum test-based Bayes factors (2) and (3). In the linear model, calibration (5) has been shown (Held & Ott, 2018) to be a lower bound on the sample-size adjusted minBF based on the g-prior for any sample size $n \geq d + 3$, where d is the number of covariates in the model (standard regularity conditions require $n \geq d + 2$). In fact, for $d = n + 3$, these minBFs converge from above to the bound (5) as $d \to \infty$, hence the subscript ∞ in the notation above.

Consider now a normally distributed test statistic $t \sim \mathrm{N}(\mu, 1)$ and the class of all prior distributions symmetric around the null value $\mu = 0$ under the alternative. To obtain the corresponding minBF for a two-sided p-value p, it suffices to consider the class of all two-point distributions symmetric around the null value. The minBF then turns out to be (Berger & Sellke, 1987; Held & Ott, 2018)

$$\mathrm{minBF}_1(p) = \min_{\mu} \frac{2\varphi(t^\star)}{\varphi(t^\star + \mu) + \varphi(t^\star - \mu)}, \tag{6}$$

where $\varphi(.)$ denotes the standard normal density function and $t^\star = \Phi^{-1}(1 - p/2) = |t|$. Since the denominator in (6) is a folded normal density function, we will refer to the minBF (6) as the "folded normal alternative" bound.

3 p-Values for 2×2 contingency tables

Next, we introduce the p-values that we are going to consider. Table 1 shows our notation for 2×2 contingency tables. To make a fair comparison between p-values and sample-size adjusted minBFs, we mainly focus on p-values which are based on the exact hypergeometric distribution of the data, avoiding large-sample approximations. We will refer to such p-values as "non-asymptotic" p-values.

	Successes ($y_i = 1$)	Failures ($y_i = 0$)	Total
Sample 1	n_{11}	n_{12}	n_{1+}
Sample 2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Table 1: Notation for 2×2 contingency tables.

However, for comparison and illustration of the convergence result (Proposition 1), we will also briefly consider the asymptotic p-value obtained from the deviance z, *i. e.* $p_{\mathrm{dev}} = 1 - F_{\chi^2(1)}(z)$, where $F_{\chi^2(1)}(.)$ denotes the cumulative distribution function of the χ^2-distribution with one degree of freedom. Note that the likelihood ratio test

underlying the deviance p-value is asymptotically equivalent to the widely used chi-squared test (Kateri, 2014). A corresponding one-sided p-value could be obtained from the signed likelihood ratio statistic (Davison, 2003, p. 128), but we will not study that p-value here.

3.1 p-Values from Fisher's exact test

While the p-value from Fisher's exact test for a one-sided significance test is uniquely defined, several different definitions of p-values have been proposed for two-sided alternatives (Fay, 2010a). We focus on three well-established definitions here, but note that even more alternative definitions have been introduced (Lloyd, 1988; Dunne et al., 1996; Meulepas, 1998).

Let N_{11} be the random variable conditional on the margins n_{1+} and n_{+1} corresponding to entry n_{11} of Table 1. The one-sided p-value for testing H_0: $\beta = 0$ against the alternative H_1: $\beta < 0$ is $P^- = \Pr(N_{11} \leq n_{11})$, where $\Pr(N_{11})$ denotes the hypergeometric distribution of N_{11} under H_0. The one-sided p-value P^+ for testing H_1: $\beta > 0$ is defined analogously.

If the alternative is two-sided, H_1: $\beta \neq 0$, then the following three definitions for a two-sided p-value have been proposed (Kateri, 2014): The "probability-based" p-value

$$p_{\text{pb}} := \sum_{t:\Pr(t)\leq\Pr(t_{\text{obs}})} \Pr(N_{11} = t),$$

the central p-value

$$p_{\text{ce}} := \min\{2\min[P^-, P^+], 1\} \tag{7}$$

and Blaker's p-value

$$p_{\text{bl}} := \min\{P^-, P^+\} + p^\star,$$

where $p^\star$ is the one-sided p-value from the other tail of the distribution than $\min\{P^-, P^+\}$,

nearest to but not larger than $\min\{P^-, P^+\}$. All these three versions of two-sided p-values can be computed using the `exact2x2()` function (Fay, 2010a,b) in the `R`-package `exact2x2` (version 1.4.1).

3.2 The mid p-value

A simple way to reduce conservativeness of the p-values from Fisher's exact test is to apply the mid-p modification, originally proposed by Lancaster (1961). The idea is to take only half of the probability mass of the observed table (and other tables that are as extreme) to compute the p-value. The mid-p modification of the one-sided p-value P^- from Fisher's exact test is $P^-_{\text{mid}} = P^- - 1/2 \cdot \Pr(N_{11} = n_{11})$ and $P^+_{\text{mid}} = 1 - P^-_{\text{mid}}$ is defined analogously. For the two-sided alternative, the mid-p modification had been applied to two versions of p-values from Fisher's exact test: Hwang & Yang (2001) suggest to use the modified probability-based p-value, where $\sum_{t:\Pr(t)=\Pr(t_{\text{obs}})} \Pr(N_{11} = t)$ is multiplied by $1/2$. Rothman & Greenland (1998, pp. 222-223) define the mid p-value in analogy to the central p-value (7):

$$p_{\text{mid}} := 2 \min\left\{P^-_{\text{mid}}, 1 - P^-_{\text{mid}}\right\}.$$

We will study this "central" mid p-value p_{mid}, which can be obtained from the function `tab2by2.test()` in the `R`-package `epitools` (version 0.5-7).

3.3 Liebermeister's significance measure

Liebermeister (1877) studied the posterior probability

$$P^-_{\text{lie}} := \Pr(\beta < 0 \,|\, n_{11}, n_{12}, n_{21}, n_{22}) \tag{8}$$

that the log odds ratio $\beta = \log\{p_1(1-p_2)/[p_2(1-p_1)]\}$ is negative (where p_i is the success probability in sample i) assuming independent uniform priors on p_1 and

p_2. He derived an exact formula for P^-_{lie} in terms of hypergeometric probabilities (Seneta, 1994). Interestingly, this posterior probability equals the one-sided p-value $P^+ = \Pr(N_{11} \geq n_{11} + 1)$ from Fisher's exact test - which has been proposed more than 50 years later - for the table obtained from Table 1 by adding one to the diagonal entries n_{11} and n_{22} (Seneta & Phipps, 2001). Overall (1980) proposed the same modification of Fisher's exact test in the frequentist context to obtain a more powerful test. Seneta & Phipps (2001) showed that Liebermeister's significance measure is less conservative than the one-sided p-value P^- from Fisher's exact test and recommended to use that significance measure instead of the p-value P^-. Thus, Liebermeister's significance measure, which is based on a default prior that does not incorporate external information, is a nice example for excellent frequentist properties of objective Bayesian procedures (Bayarri & Berger, 2004). Altham (1969) obtained a formula for the posterior probability in (8) under the more general assumption of two independent beta distributions on the two success probabilities and established a relation to Fisher's exact test in this setting. Nurminen & Mutanen (1987) further generalized these results and derived a formula for the whole posterior distribution of the odds ratio, see Ly (2017, chapter 11) for an extension to localized beta priors. Jeffreys' prior and dependent prior distributions for the two success probabilities are considered in Howard (1998).

To obtain a two-sided significance measure from (8), we apply the same strategy as for the central p-value in Fisher's exact test (this strategy is also mentioned in Overall, 1980):

$$p_{\text{lie}} := 2 \min \left\{ P^-_{\text{lie}}, 1 - P^-_{\text{lie}} \right\}.$$

The probabilities P^-_{lie} and p_{lie} will be called (one- and two-sided) Liebermeister's p-value in the sequel.

3.4 Application

We now illustrate the different types of (two-sided) p-values using the dataset on appendicitis patients presented in Table 2 (data taken from Seneta & Phipps (2001), originally reported in Di Sebastiano et al. (1999)). The aim was to investigate the relationship between severe right abdominal pain and acute inflammation of the appendix.

	Pain	No pain	Total
Acute	1	15	16
Non-acute	5	10	15
Total	6	25	31

Table 2: Observed number of acute and non-acute appendicitis patients with and without severe right abdominal pain in the study by Di Sebastiano et al. (1999).

The conditional maximum likelihood estimate (Breslow, 1981) of the odds ratio for Table 2 is 0.142 (95 % CI from 0.003 to 1.542). There is substantial variation in the p-values for that table, ranging from the deviance p-value $p_{\text{dev}} = 0.049$ to the central p-value $p_{\text{ce}} = 0.14$, see Table 3 for the other types of p-values. So there is some evidence against H_0 for p_{dev}, only weak evidence against H_0 for p_{lie}, p_{mid} and p_{pb}, but little or no evidence for p_{ce} - which is about twice as large as p_{lie} - according to Bland (2015, p. 117). Given this variation in p-values, one may wonder which one should be used (assuming one wants to use a p-value at all). We will provide suggestions from a Bayesian viewpoint later. A first step to a Bayesian measure of evidence would be to compute one of the large-sample minBFs (2), (3) or (6) for two-sided p-values. These minBFs for Table 2 obtained from the five different types of p-values are shown in Table 3. However, we will see later that these bounds are suboptimal for small and moderate sample sizes.

	p-value	minTBF_1	minTBF_2	minBF_1
Central	0.144	1/1.2	1/1.3	1/1.5
Prob.-based	0.083	1/1.6	1/1.8	1/2.3
Mid P	0.079	1/1.6	1/1.8	1/2.3
Liebermeister	0.069	1/1.7	1/2	1/2.6
Deviance	0.049	1/2.1	1/2.5	1/3.5

Table 3: Non-asymptotic p-values, the asymptotic deviance p-value and the corresponding large-sample minBFs (2), (3) and (6) for Table 2.

4 Sample-size adjusted minimum Bayes factors

In this section, we introduce a sample-size adjusted minBF for GLMs over a class of (multivariate) normal priors. First, consider the linear model

$$\boldsymbol{y} = \alpha\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\boldsymbol{I}), \tag{9}$$

where α is the intercept, $\boldsymbol{X}$ the design matrix, $\boldsymbol{\beta}$ the vector of regression coefficients and σ^2 the residual variance. To obtain the Bayes factor for the null model (with intercept α only) against the linear model (9), the g-prior $\boldsymbol{\beta} \,|\, g, \sigma^2 \sim \text{N}(\mathbf{0}, g\,\sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ (Zellner, 1986) is assigned to the vector of regression coefficients and a reference prior $f(\alpha, \sigma^2) \propto \sigma^{-2}$ to the intercept α and the residual variance σ^2. Under these priors, there is a closed-form expression for the Bayes factor (Liang et al., 2008), which is minimized if the prior variance factor g is estimated by empirical Bayes. This yields a closed-form expression for the minBF, which depends only on the sample size n, the dimension d of $\boldsymbol{\beta}$ and the coefficient of determination R^2 (Held & Ott, 2016, equation (9)). A p-value from the F-test, a one-to-one transformation of R^2, can hence be mapped to this minBF, see figure 1 in Held & Ott (2016) for illustration. In the case $d = 2$, there is even a closed-form expression for the minBF as a function of the p-value and the sample size (Held & Ott, 2016, equation (12)). For the other dimensions d, the conversion from the F-test p-value to R^2 has to be done numerically.

4.1 Generalized linear model

At least two difficulties arise if we want to extend the procedure described above to a GLM: First, several different definitions of generalized g-priors for GLMs have been proposed (see *e. g.* Hansen & Yu, 2003; Sabanés Bové & Held, 2011; Li & Clyde, 2016) and second, there is no closed-form expression for the marginal likelihood under most of these priors, making computation of the Bayes factor more involved. An exception is the proposal by Li & Clyde (2016), where a closed-form expression for the approximate marginal likelihood is available. We opted for this approach, mainly due to computational efficiency, but there is also a theoretical motivation that will be presented in Section 4.2. Furthermore, it is not possible to obtain a direct mapping between p-values and minBFs for discrete data such as 2×2 contingency tables. We will therefore study the relationship between p-values and minBFs by fitting a regression model to all observed pairs of p-value and minBF (one for each 2×2 table) for fixed sample size n.

Let $\boldsymbol{X}$ denote the design matrix of a GLM having the covariate vectors $\boldsymbol{x}_k$, $k = 1, \ldots, d$, as columns, $\boldsymbol{\beta}$ the vector of regression coefficients of dimension d and α_i, $i = 1, 2$, intercept terms. We compare the null model $\mathcal{M}_0$ with linear predictor vector $\boldsymbol{\eta}_0 = \alpha_0 \mathbf{1}$ to model $\mathcal{M}_1$ with linear predictor $\boldsymbol{\eta}_1 = \alpha_1 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}$. Before defining the prior, Li & Clyde (2016) apply a centering step to make the intercept and the regression coefficients locally orthogonal. This leads to the reparametrization $\boldsymbol{\eta}_1 = \alpha_1^c \mathbf{1} + \boldsymbol{X}^c \boldsymbol{\beta}$ of model $\mathcal{M}_1$, where $\boldsymbol{X}^c$ denotes the orthogonalized design matrix and α_1^c the shifted intercept (see Li & Clyde, 2016, section 2.2 for details). After this reparametrization, the observed Fisher information matrix is block-diagonal with the observed Fisher information (matrices) $\mathcal{J}_n(\hat{\alpha}_1^c)$ and $\mathcal{J}_n(\hat{\boldsymbol{\beta}})$ as blocks. Now, independent priors will be assigned to the regression coefficients and the intercept: The generalized g-prior (for

$g \geq 0$)

$$\boldsymbol{\beta} \,|\, g, \mathcal{M}_1 \sim \mathrm{N}\left(\mathbf{0}, g\mathcal{J}_n\left(\hat{\boldsymbol{\beta}}\right)^{-1}\right), \tag{10}$$

usually accompanied with an improper prior $p(\alpha_1^c) \propto 1$. The prior for $\boldsymbol{\beta}$ depends on the sample size and the data, in particular also on the outcome $\boldsymbol{y}$, through $\mathcal{J}_n(\hat{\boldsymbol{\beta}})$.

A useful property of prior (10) is its invariance with respect to location-scale transformations of the covariates (Li & Clyde, 2016, section 4.2), which is one of the desiderata for model selection priors (Bayarri et al., 2012). An additional important advantage of prior (10) is computational efficiency, since an approximate marginal likelihood (see Li & Clyde (2016), equation (17)) as a function of the maximum likelihood estimates (MLEs) can be obtained in closed form by applying an integrated Laplace approximation (Wang & George, 2007).

For fixed g, the Bayes factor for the null model $\mathcal{M}_0$ (with an improper prior $p(\alpha_0) \propto 1$) against model $\mathcal{M}_1$ then turns out to be

$$\mathrm{BF}(g) \approx \exp\left(-\frac{z}{2}\right)\left[\frac{\mathcal{J}_n\left(\hat{\alpha}_1^c\right)}{\mathcal{J}_n\left(\hat{\alpha}_0\right)}\right]^{1/2}(1+g)^{d/2}\exp\left[\frac{Q}{2(1+g)}\right] \tag{11}$$

(Li & Clyde (2016), equation (19)), where z denotes the deviance statistic $z(\boldsymbol{y}) = 2\log\left[f(\boldsymbol{y} \,|\, \hat{\alpha}_1, \hat{\boldsymbol{\beta}}, \mathcal{M}_1)/f(\boldsymbol{y} \,|\, \hat{\alpha}_0, \mathcal{M}_0)\right]$ and

$$Q := \hat{\boldsymbol{\beta}}^\top \mathcal{J}_n(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}$$

is the squared Wald statistic. (Note that the centering step does not change the deviance statistic z nor the squared Wald statistic Q.) The Bayes factor (11) is minimized if we estimate the parameter g such that the marginal likelihood under $\mathcal{M}_1$ is maxi-

mized, *i. e.* if we estimate g by empirical Bayes, which yields

$$\hat{g} = \max\{Q/d - 1, 0\} \qquad \text{and} \tag{12}$$

$$\text{minBF} = \text{BF}(\hat{g}). \tag{13}$$

4.2 Asymptotic behavior

We are also interested in the large-sample and asymptotic properties of the sample-size adjusted calibrations we will develop. To have a reference line for these calibrations, we will now study convergence of the approximate minBF (13). To do so, we consider a so-called sequence of local alternative hypotheses $(H_1^n)_{n\in\mathbb{N}}$, where H_1^n: $\boldsymbol{\beta} = \mathcal{O}(n^{-1/2})$ (Davidson & Lever, 1970), so the true regression coefficients get smaller with increasing sample size n. Johnson (2005, p. 691) argues that for moderately large sample sizes, this is the case of practical interest, since for larger $\boldsymbol{\beta}$, it would be trivial to differentiate between H_0 and H_1^n, whereas for smaller $\boldsymbol{\beta}$, it would be too difficult. Under such a sequence of local alternatives (and some regularity conditions), the distribution of the deviance converges to a non-central χ^2-distribution, which is used in the derivation of the minimum test-based Bayes factor (1) (Johnson, 2008). Motivated by the asymptotic equivalence of the squared Wald statistic Q and the deviance z under a sequence of local alternatives and under the null hypothesis, we now establish a convergence result for fixed deviance z:

Proposition 1. *If* $\lim_{n\to\infty} Q = z$ *and* $\lim_{n\to\infty} \mathcal{J}_n\left(\hat{\boldsymbol{\alpha}}_1^c\right) / \mathcal{J}_n\left(\hat{\boldsymbol{\alpha}}_0\right) = 1$, *then*

$$\lim_{n\to\infty} \textit{minBF} = \textit{minTBF}_d,$$

where minBF is the minimum Bayes factor (13) *(bounded by* 1 *from above) for sample size* n *and* $\textit{minTBF}_d$ *the minimum test-based Bayes factor* (1).

Note that Li & Clyde (2016, section 2.4) obtained a closely related result for the Bayes

factor (11) with fixed g. The proof of Proposition 1 can be found in the Supporting Information, Section 1.1. In fact, it can be shown that under the null hypothesis H_0: $\boldsymbol{\beta} = \mathbf{0}$ and under a sequence of local alternative hypotheses H_1^n: $\boldsymbol{\beta} = \mathcal{O}(n^{-1/2})$ (and some common regularity conditions), the two conditions of Proposition 1 are satisfied for a GLM with canonical link function in a weaker sense, *i.e.* $Q - z \xrightarrow{P} 0$ (Engle, 1984, also holds for non-canonical link function) and $\mathcal{J}_n(\hat{\alpha}_1^c)/\mathcal{J}_n(\hat{\alpha}_0) \xrightarrow{P} 1$ (proof in the Supporting Information, Section 1.2). Here, $\xrightarrow{P}$ denotes convergence in probability for $n \to \infty$. Since there is a continuous one-to-one mapping between the deviance z and the deviance p-value p_{dev}, we also have $\text{minBF}(p_{\text{dev}}) \to \text{minTBF}_d(p_{\text{dev}})$ as $n \to \infty$ for fixed deviance p-value p_{dev} and any GLM with canonical link function. See Figure 2 for an illustration of this result in the logistic regression setting with $d = 1$.

5 Sample-size adjusted minimum Bayes factors for 2×2 tables

We will focus on two classes of prior distributions for the log odds ratio β under the two-sided alternative H_1: $\beta \neq 0$. A common choice are local priors, *i.e.* unimodal distributions which are symmetric around 0. The main motivation for choosing a class of normal priors among the local priors is that these priors are invariant with respect to location-scale transformations of the covariate. Moreover, minBFs based on normal priors allow for comparison with the sample-size adjusted minBFs based on the g-prior in the linear model, where that prior is a natural choice. The second class, the class of all prior distributions symmetric around $\beta = 0$, can be claimed to contain all "reasonable" priors. We will see that the minimum of the Bayes factor over these symmetric priors is actually attained at a symmetric two-point distribution (including the degenerate case $\beta = 0$). This corresponds to the two-sided version of a simple alternative. In the one-sided case, we will restrict attention to simple alternatives.

5.1 Local normal alternatives

Consider the logistic regression model $\mathcal{M}_1$ with Bernoulli outcomes $y_i \in \{0, 1\}$, logistic link function $\text{logit}\{\Pr(y_i = 1)\} = \eta_{1,i}$ and linear predictor vector $\boldsymbol{\eta}_1 = \alpha_1 \mathbf{1} + \beta \boldsymbol{x}$. Here, $\boldsymbol{x}$ is a vector of binary covariate values, taking the value 1 for an observation from sample 1 and 0 otherwise. The null model $\mathcal{M}_0$ corresponding to the null hypothesis H_0: $\beta = 0$ has the linear predictor vector $\boldsymbol{\eta}_0 = \alpha_0 \mathbf{1}$. We assume improper uniform priors on the intercepts α_0 and α_1^c.

Our goal is to obtain a lower bound on the (approximate) Bayes factor (11) for any generalized g-prior (10) on β. Note that prior (10) here boils down to a univariate normal prior centered at zero (as used in Edwards et al. (1963) for the normal likelihood). As explained in the previous section, we will fix the parameter g at the empirical Bayes estimate (12).

We exclude 2×2 tables with entries equal to zero from our analysis, as the MLE $\hat{\beta}$ (and possibly also $\hat{\alpha}_0$ and/or $\hat{\alpha}_1$) does not exist for such tables. For the remaining tables, there are closed-form expressions for the MLEs of α_0, α_1 and β, the corresponding observed Fisher information $\mathcal{J}_n(\hat{\alpha}_0)$, $\mathcal{J}_n(\hat{\alpha}_1)$ and $\mathcal{J}_n(\hat{\beta})$, as well as the deviance z. Consequently, the approximate minBF (13) can in principle be written as a closed-form expression that depends on the four entries n_{11}, n_{12}, n_{21} and n_{22} of Table 1 only, but we will present a more concise representation that is easier to interpret.

To this end, note that $\hat{\beta} = \log(n_{11}) - \log(n_{12}) - \log(n_{21}) + \log(n_{22})$ is the MLE of the log odds ratio and

$$\mathcal{J}_n(\hat{\alpha}_0) = \left(n_{+1}^{-1} + n_{+2}^{-1}\right)^{-1}, \tag{14}$$

$$\mathcal{J}_n(\hat{\alpha}_1^c) = \left(n_{11}^{-1} + n_{12}^{-1}\right)^{-1} + \left(n_{21}^{-1} + n_{22}^{-1}\right)^{-1}, \tag{15}$$

$$\mathcal{J}_n(\hat{\beta}) = \left(n_{11}^{-1} + n_{12}^{-1} + n_{21}^{-1} + n_{22}^{-1}\right)^{-1}, \tag{16}$$

as derived in the Supporting Information (Section 2). In that supplementary section,

we also present a closed-form expression for the deviance z, which is equivalent to

$$\exp(-z/2) = n_{11}^{-n_{11}} n_{12}^{-n_{12}} n_{21}^{-n_{21}} n_{22}^{-n_{22}} n_{1+}^{n_{1+}} n_{2+}^{n_{2+}} n_{+1}^{n_{+1}} n_{+2}^{n_{+2}} n^{-n}. \tag{17}$$

Now, to determine the minBF, the first step is to compute the squared Wald statistic $Q = \hat{\beta}^2 \mathcal{J}_n(\hat{\beta})$ and then to infer $\hat{g} = \max\{Q - 1, 0\}$. The approximate minBF is then given by

$$\text{minBF} \approx \min \left\{ \left[\frac{\mathcal{J}_n(\hat{\alpha}_1^c)}{\mathcal{J}_n(\hat{\alpha}_0)} (1 + \hat{g}) \right]^{1/2} \exp \left[\frac{Q}{2(1+\hat{g})} - \frac{z}{2} \right], 1 \right\}. \tag{18}$$

Expression (18) will be called the "local alternative" minBF. If $Q > 1$, so that $\hat{g} > 0$, then

$$\text{minBF} \approx \min \left\{ \left[\frac{\mathcal{J}_n(\hat{\alpha}_1^c)}{\mathcal{J}_n(\hat{\alpha}_0)} \right]^{1/2} (e \cdot Q)^{1/2} \exp \left(-\frac{z}{2} \right), 1 \right\}. \tag{19}$$

Kass & Vaidyanathan (1992, theorem 3) have obtained a closely related, but more complex approximate formula than (19) for the minBF over local normal priors. Since these authors assume the parameters α_1 and β to be null orthogonal (*i. e.* the expected Fisher information matrix is block-diagonal for $\beta = 0$ and all α_1) instead of locally orthogonal at the MLE, the observed Fisher information matrix is not block-diagonal. Thus, the factor $\mathcal{J}_n(\hat{\alpha}_1^c)\mathcal{J}_n(\hat{\beta})$ of the term $\mathcal{J}_n(\hat{\alpha}_1^c)Q$ in Equation (19) is replaced by the determinant of the observed Fisher information $\mathcal{J}_n(\hat{\alpha}_1, \hat{\beta})$ (in the null orthogonal parametrization), which is a more complex expression than the product of (15) and (16). Furthermore, the approximation by Kass & Vaidyanathan (1992) is valid only if the true value of β lies within an $\mathcal{O}(n^{-1/2})$ neighborhood of the null value $\beta = 0$. This condition does not hold for unbalanced 2×2 tables with large log odds ratios β, which have small associated p-values. Since we are mainly interested in 2×2 tables with small p-values, we prefer the Li & Clyde (2016) approach, where no such restriction is

imposed.

Note that the minBF (18) is invariant with respect to row-column switch in the 2×2 table (*i. e.* exchanging the two entries n_{12} and n_{21}), which implies that exchanging the covariate $\boldsymbol{x}$ and the outcome $\boldsymbol{y}$ will not affect the minBF. This is useful for retrospective epidemiological studies, for example, where either the case/control status (prospective model) or the exposure variable (retrospective model) may be taken as outcome (Breslow & Powers, 1978). That invariance property is not hard to establish since the squared Wald statistic Q, the deviance z and the ratio $\mathcal{J}_n(\hat{\alpha}_1^c)/\mathcal{J}_n(\hat{\alpha}_0)$ all turn out to be invariant under row-column switch.

The minBF (18) for Table 2 is $1/2.6$. So a lower bound on $\Pr(H_0 \,|\, \text{data})$ is $0.38/1.38 = 0.28$ under the assumption of equipoise, *i. e.* $\Pr(H_0) = \Pr(H_1) = 0.5$ (Johnson, 2013). Note that both large-sample minBFs (2) and (3) given in Table 3 are larger than the sample-size adjusted minBF (18) for all types of p-values, the differences between these bounds and (18) being rather large for all types of non-asymptotic p-values. This illustrates that applying the large-sample minBFs (2) and (3) to non-asymptotic p-values from 2×2 tables of small or moderate sample size may lead to substantial overestimation of the minBF.

5.2 Two-sided simple alternatives

Here, we aim at finding a discrete analogon of the "folded normal alternative" minBF (6). To obtain the minBF over the class of all prior distributions on the log odds ratio β symmetric around 0, it suffices to consider the subclass of all symmetric two-point distributions (Berger & Sellke, 1987). For such two-point distributions, the marginal likelihood under the alternative H_1 is of the form

$$\frac{1}{2}\left\{f(n_{11}; \exp[-\beta]) + f(n_{11}; \exp[\beta])\right\},$$

where $f(n_{11};\theta)$ denotes the probability mass function of the non-central hypergeometric distribution with non-centrality parameter θ. The likelihood under H_0 is then $f(n_{11};1)$. Thus, ignoring the fact that the log odds ratio β can only take values in a discrete set, a minBF over the class of symmetric priors for a 2×2 table with non-zero entries can be defined as

$$\text{minBF} = \min_{\beta} \frac{2f(n_{11};1)}{f(n_{11};\exp[-\beta]) + f(n_{11};\exp[\beta])}. \tag{20}$$

Note that the denominator of (20) is an analogon of the folded normal density in the "folded normal alternative" bound (6) and could thus be called "folded hypergeometric density". The bound (20) will be termed "two-sided simple alternative" minBF in the sequel. This minBF is not larger than the local alternative minBF (18) since the class of local normal priors is contained in the class of prior distributions symmetric around the null value.

The two-sided simple alternative minBF (20) for Table 2 is 1/3.3, corresponding to a lower bound of 0.23 on $\Pr(H_0 \mid \text{data})$ under the assumption of equipoise. For all types of non-asymptotic p-values, the minBF (20) is smaller than the large-sample "folded normal alternative" bound (6), see Table 3.

5.3 One-sided simple alternatives

For the one-sided alternative H_1: $\beta < 0$, the minBF over the class of simple alternatives in the direction of H_1 is given by

$$\text{minBF} = \min_{\beta \leq 0} \frac{f(n_{11};1)}{f(n_{11};\exp[\beta])}. \tag{21}$$

Expression (21) will be called ''one-sided simple alternative" minBF. Note that this minBF never exceeds the two-sided simple alternative minBF (20) and is usually about half as large as minBF (20) given the latter is not close to 1.

6 Empirical relationship between p-values and minimum Bayes factors

To investigate how the maximal evidence of p-values for 2×2 tables against the point null hypothesis depends on the sample size, we will study the relationship between the p-values introduced in Section 3 and the sample-size adjusted minBFs (18), (20) and (21) using a novel parametric regression approach. Table 4 gives an overview of the types of p-values considered.

p-value	One-sided, H_1: $\beta < 0$	Two-sided, H_1: $\beta \neq 0$
Fisher	-	Prob.-based p_{pb}
	P^-	Central $p_{\text{ce}} = \min\{2\min[P^-, P^+], 1\}$
	-	Blaker p_{bl}
Mid p	P^-_{mid}	$p_{\text{mid}} = 2\min\left\{P^-_{\text{mid}}, 1 - P^-_{\text{mid}}\right\}$
Liebermeister	P^-_{lie}	$p_{\text{lie}} = 2\min\left\{P^-_{\text{lie}}, 1 - P^-_{\text{lie}}\right\}$
Asymptotic	-	Deviance p_{dev}

Table 4: The different types of p-values considered for testing H_0: $\beta = 0$ against either a one-sided (left column) or two-sided (right column) alternative H_1. See Section 3 for the definitions of the p-values. The p-values in the left column will be related to the one-sided simple alternative minBF (21), whereas the p-values in the right column will be related to the both the local normal alternative minBF (18) and two-sided simple alternative minBF (20).

6.1 Study description

Suppose we are interested in the relationship between the probability-based p-value and the local alternative minBF (18), for example. Then, there is a p-value p and a sample-size adjusted minimum Bayes factor minBF associated to each 2×2 table. For fixed n, we consider all such pairs (p, minBF) for 2×2 tables with non-zero entries of sample size n and fit a specific regression model to these points, as illustrated in

Figure 1 for $n = 20$. The number of different tables for sample size n is $\binom{n-1}{3}$, which corresponds to 969 tables for $n = 20$, 3'654 for $n = 30$, 18'424 for $n = 50$, 156'849 for $n = 100$, 4'410'549 for $n = 300$ and 165'668'499 tables for $n = 1000$. Due to restrictions in memory and computation time, for sample size $n = 1000$, we computed the p-values and minBFs for a random sample of size 10^6 from all possible 2×2 tables only. All the following analyses for $n = 1000$ are based on such a random sample. As the sample size n increases, the first to third quartiles of the distribution of the observed p-values and minBFs decrease, as well as the minimum of these quantities. For example, for $n = 50$, the minimum observed probability-based p-value is $p_{\text{pb}} = 5.8 \times 10^{-12}$, for $n = 100$, it is $p_{\text{pb}} = 2.7 \times 10^{-26}$, and for $n = 1000$, it is around $p_{\text{pb}} = 10^{-292}$.

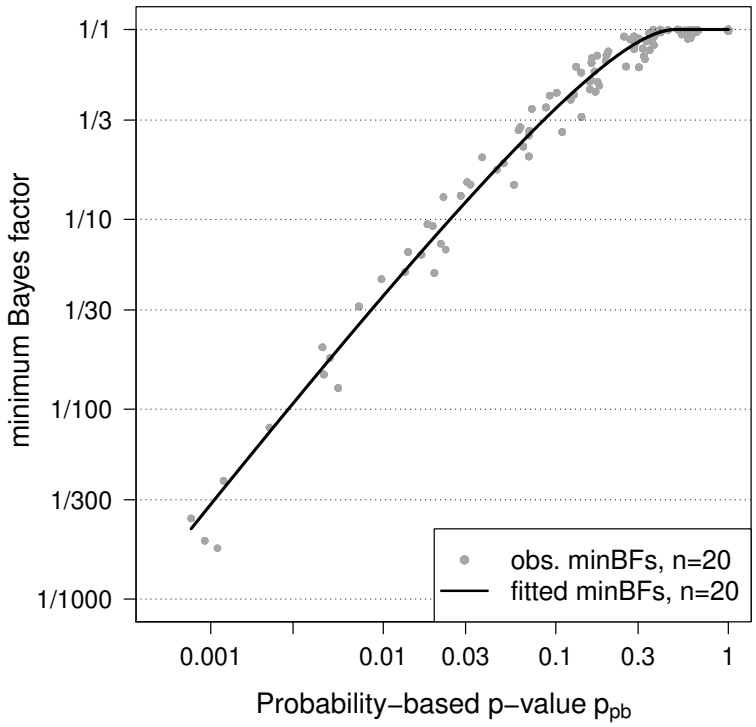

Figure 1: The points represent the pairs of probability-based p-values and local alternative minBFs (18) that are observed for 2×2 tables (with non-zero entries) of sample size $n = 20$. Fitting model (22) to these points then leads to the fitted minBFs for $n = 20$ shown as a line.

We consider two parametric candidate models. These regression models are motivated by the large-sample "$-ep\log(p)$" calibration (3) or the "$-eq\log(q)$" calibration (5), respectively, which directly relate two-sided p-values to minBFs. For example, Equation (3) implies

$$\log(\text{minTBF}_2) = \begin{cases} 1 + \log(p) + \log[-\log(p)] & \text{for } p < 1/e, \\ 0 & \text{otherwise.} \end{cases}$$

We will instead fit a model of the form

$$\log(\text{minBF}) = \begin{cases} a + b_1 \log(p) + b_2 \log[-\log(p)] & \text{for } p < \exp(-b_2/b_1) \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

with least squares to all pairs (p, minBF) (for fixed n) under the constraint $a = b_2[1 - \log(b_2/b_1)]$. This formulation ensures that the resulting fitted curve is monotonically increasing and continuous, *i.e.* the non-constant part of (22) has its maximum at 0 (provided $b_2 > 0$). For the "$-eq\log(q)$" calibration (5), the corresponding model is

$$\log(\text{minBF}) = \begin{cases} a + b_1 \log(1-p) + b_2 \log[-\log(1-p)] & \text{for } p < 1 - \exp(-b_2/b_1) \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

and the constraint on the intercept a is the same as above. For fixed n, we fit both model (22) and (23) and quantify the goodness of fit by R^2. For example, the fit shown in Figure 1 is based on model (22) with $a = 1.19$, $b_1 = 1.25$ and $b_2 = 0.88$ and $R^2 = 0.97$. For each class of alternatives studied, we compare the values of R^2 for model (22) and (23) and then choose the model with the larger values for the majority of cases (pairs of type of p-value and sample size) considered. Note that in model (22), the parameter b_1 corresponds to the asymptotic slope of the curve for $\log(p) \to -\infty$, whereas in model (23), the asymptotic slope is b_2. This slope describes the (linear) relationship

between $\log(\text{minBF})$ and $\log(p)$ for small p.

Observe that most pairs (p, minBF) occur multiple times as they correspond to different 2×2 tables. The models (22) and (23) are fitted to all these pairs including multiple pairs. However, pairs (p, minBF) with p-value $p < 10^{-6}$ are ignored to ensure a good fit in the practically interesting range for large sample sizes. The marginal distributions of the p-values restricted to the range $[10^{-6}, 1]$ and the corresponding local alternative minBFs are illustrated in Figure 2. For the one-sided p-values, we restrict our analyses to 2×2 tables with log odds ratio $\beta \leq 0$ in the direction of the alternative H_1: $\beta < 0$.

6.2 Results for local normal alternatives

For the local alternative minBF (18), R^2 is consistently higher for model (22) than for model (23) for all types of p-values and sample sizes considered. Despite relatively small differences in R^2 (lying in the interval $[0.002, 0.009]$), discrepancies between the fitted curves for the two models are visible, especially for larger p-values in the range $[0.03, 0.3]$. It is plausible that model (22) fits better, since the "$-ep\log(p)$" calibration (3) underlying this model is also based on a generalized g-prior on the vector of regression coefficients. So we choose model (22) for all analyses of local alternative minBFs.

We have also approximated the logarithm of the large-sample bound minTBF_1 (2) by model (22). To do so, we performed least squares function approximation (*i. e.* minimizing the integral over the squared differences between the two functions). The estimated coefficients turned out to be $\hat{a} = 0.99671$, $\hat{b}_1 = 1.00770$ and $\hat{b}_2 = 1.16017$. Goodness of fit is excellent with a maximum absolute difference between the fitted curve (22) and $\log(\text{minTBF}_1)$ of 0.0023 for p-values $p \in [10^{-6}, 1]$, see Figure S1 in the Supporting Information for a plot of this difference as a function of the p-value. Hence, for p-values in this range, the bound $\text{minTBF}_1(p)$ - which needs to be computed

numerically - can be very well approximated by the simple parametric model (22) with the coefficients given above. These coefficients can then be compared with the corresponding coefficients obtained by fitting model (22) to the pairs of p-values and local alternative minBFs for different sample sizes n.

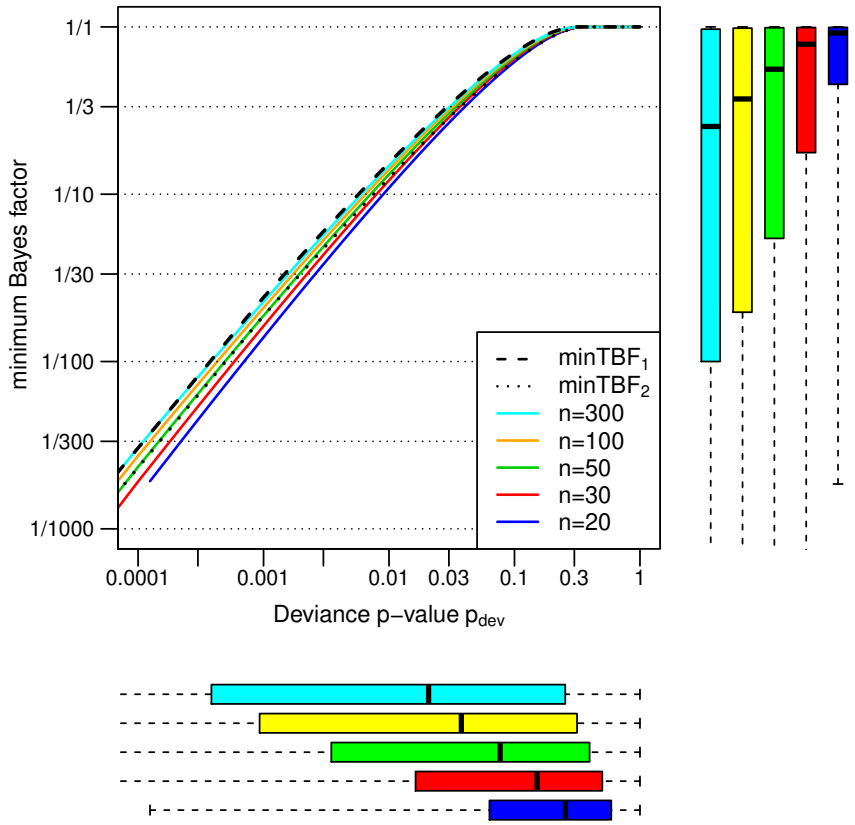

Figure 2: The colored curves show the estimated mean relationship (22) between the deviance p-values and the local alternative minBFs (18) for sample size $n = 20, 30, 50, 100, 300$. The broken lines represent the large-sample minBFs (2) and (3). The marginal boxplots illustrate the distribution of the observed deviance p-values restricted to the range $[10^{-6}, 1]$ and the corresponding minBFs for the same set of sample sizes. For the other types of p-values, the boxplots below the x-axis look similar, although the range and position of the boxes varies slightly.

The fitted local alternative minBFs based on the deviance p-value from model (22)

are displayed in Figure 2 and the corresponding estimated coefficients are given in Table 5. Goodness of fit is excellent ($R^2 > 0.996$ for all sample sizes considered). As expected from Proposition 1, for fixed deviance p-value p_{dev}, these fitted minBFs tend towards the large-sample bound $\text{minTBF}_1(p_{\text{dev}})$ as the sample size goes to infinity. Next, we will study the same relationship between p-values and fitted minBFs for the non-asymptotic p-values in more detail.

p-value	**Prob.-based**			**Central**			**Blaker**		
n	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$
20	1.193	1.254	0.885	0.971	1.204	0.537	1.168	1.238	0.845
30	1.163	1.189	0.948	1.033	1.167	0.654	1.161	1.188	0.944
50	1.094	1.107	0.942	1.023	1.102	0.712	1.091	1.105	0.934
100	1.055	1.058	0.974	1.022	1.057	0.795	1.053	1.057	0.968
300	1.028	1.028	1.037	1.026	1.032	0.926	1.028	1.028	1.034
1000	1.015	1.017	1.091	1.021	1.021	1.027	1.015	1.017	1.090

p-value	**Mid** p			**Liebermeister**			**Deviance**		
n	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$
20	1.292	1.307	1.114	1.280	1.280	1.305	1.093	1.093	1.131
30	1.220	1.223	1.133	1.196	1.198	1.280	1.071	1.074	1.153
50	1.134	1.134	1.101	1.108	1.111	1.191	1.041	1.046	1.142
100	1.077	1.078	1.113	1.053	1.057	1.149	1.021	1.027	1.144
300	1.033	1.038	1.141	1.020	1.027	1.147	1.007	1.016	1.152
1000	1.011	1.020	1.156	1.006	1.015	1.154	1.001	1.011	1.158

Table 5: The estimated coefficients in model (22) for different sample sizes n, using the probability-based, central, Blaker's, mid p, Liebermeister's or deviance p-value, respectively.

Figure 3 shows the fitted local alternative minBFs for different sample sizes and types of non-asymptotic p-values. For all of these p-values, the fitted minBFs increase as the sample size increases, which indicates that the maximal evidence of these p-values against H_0 is on average inversely related to sample size. The coefficient estimates $\hat{a}$, $\hat{b}_1$ and $\hat{b}_2$ in model (22) for the different types of p-values are presented in Table 5. For Blaker's p-value, these estimates and hence also the fitted minBFs are very similar to the ones for the probability-based p-value (see Table 5 and Figure S2 in the

Supporting Information). In fact, the probability-based p-value and Blaker's p-value are equal or very similar for the majority of the 2×2 tables considered.

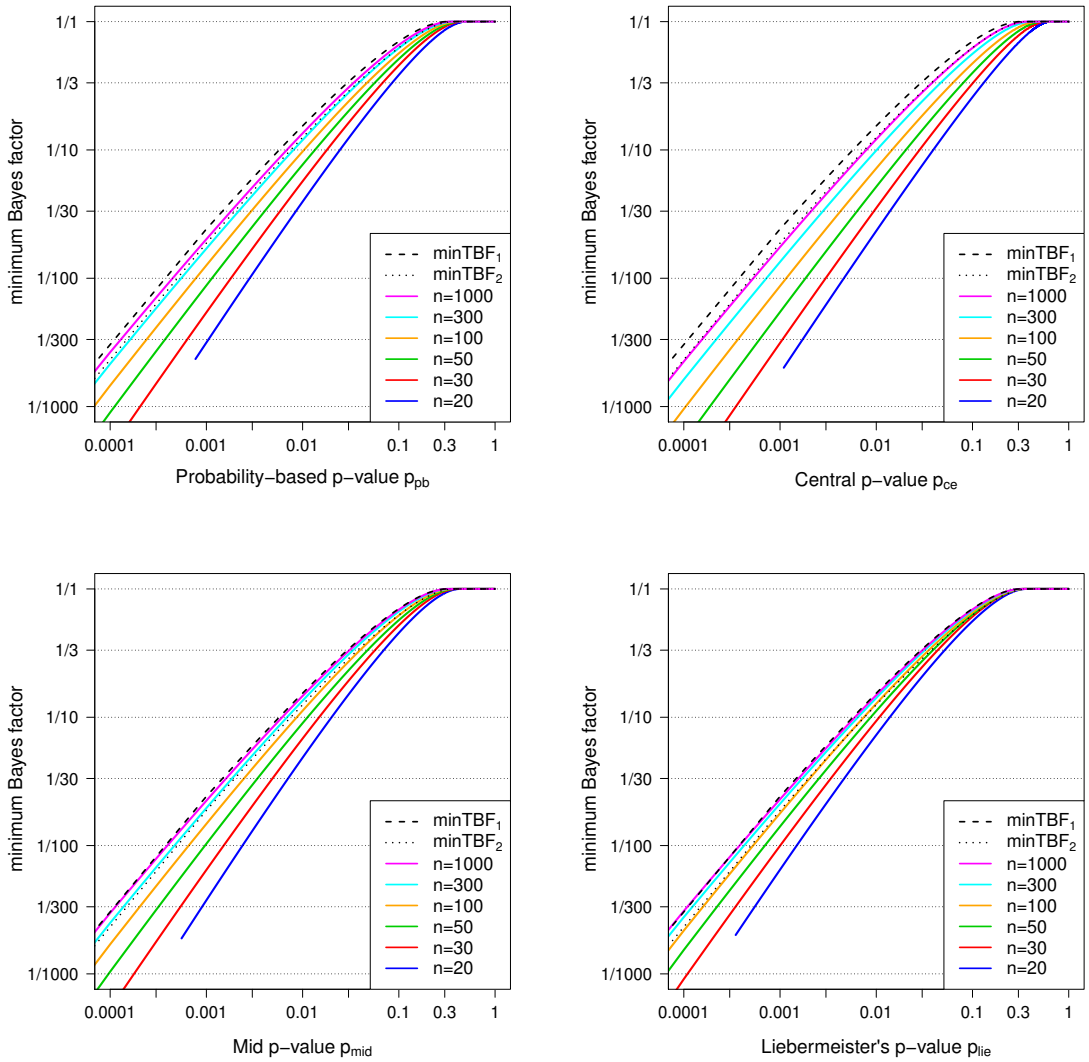

Figure 3: The colored curves show the estimated mean relationship (22) between two-sided p-values and local alternative minBFs (18) for sample size $n = 20, 30, 50, 100, 300, 1000$, using the probability-based p-value (top left), the central p-value (top right), the mid p-value (bottom left) and Liebermeister's p-value (bottom right). For comparison, the large-sample bounds (2) and (3) are also displayed.

It turns out that model (22) fits very well, achieving values of $R^2 > 0.96$ for all

types of p-values and all sample sizes considered (Table S1 in the Supporting Information). For small or moderate sample size $n \in \{20, 30\}$, the goodness of fit is best for the mid p-value, followed by the central p-value and then Liebermeister's p-value and it is worst for the probability-based p-value. For larger sample size $n \in \{50, 100, 300, 1000\}$, the ranking in terms of goodness of fit is the same except that the ranks of the probability-based p-value and Liebermeister's p-value are swapped.

Due to convergence of the sample-size adjusted local alternative minBFs (18) to the bound minTBF_1 (2) for fixed deviance p-value (Proposition 1), it is interesting to study the large-sample behavior of the fitted minBFs for the non-asymptotic p-values with respect to that bound. Figure 3 indicates that convergence to minTBF_1 occurs for the fitted minBFs based on Liebermeister's p-value and the mid p-value, but not necessarily (or only very slowly) for the standard p-values (probability-based and central p-value) from Fisher's exact test. Accordingly, for $n = 1000$, the estimated coefficients in the fitted model (22) are very similar to the estimates for the bound minTBF_1 (given at the beginning of Section 6.2) for Liebermeister's and the mid p-value, but not for the other p-values. Note also that the estimated asymptotic slopes $\hat{b}_1$ in Table 5 decrease monotonically towards 1.008 - the corresponding estimate for the large-sample bound minTBF_1 - for all p-values as the sample size n increases. These coefficient estimates $\hat{b}_1$ are especially relevant, since they mainly account for the differences in curves observed for different sample sizes n and the same type of p-value. The coefficient estimate $\hat{b}_2$ determines the p-value threshold at which the corresponding fitted minBFs reach 1 (their maximum). These estimates $\hat{b}_2$ vary most for the most conservative central p-value.

6.3 Results for two-sided simple alternatives

For the two-sided simple alternative minBF (20), R^2 is consistently higher for model (23) than for model (22) for all types of p-values and sample sizes considered (differ-

ences in R^2 are in $[0.001, 0.007]$). Thus, we use model (23) in all analyses of two-sided simple alternative minBFs. As expected, for fixed deviance p-value, the fitted two-sided simple alternative minBFs tend towards the corresponding large-sample "folded normal alternative" bound (6) as the sample size goes to infinity (see the left panel of Figure S3 in the Supporting Information).

We now turn to the non-asymptotic p-values. For the more general class of prior distributions symmetric around the null value $\beta = 0$, we observe the same qualitative relationship between the maximal evidence of the these p-values against H_0 and sample size as for the class of local normal priors, see Figure 4. However, the minBFs over this larger class depend on average less strongly on sample size than the local alternative minBFs. Goodness of fit is also excellent for the two-sided simple alternative minBFs with $R^2 > 0.97$ for all sample sizes and types of p-values considered (Table S3 in the Supporting Information). The ranking in terms of goodness of fit is the same as for the local alternative minBFs, except that the swap in ranks between Liebermeister's p-value and probability-based p-value occurs at larger sample size (between $n = 50$ and $n = 100$) for the two-sided simple alternative minBFs.

Due to asymptotic normality of the log odds ratio, we expect convergence of the fitted two-sided simple alternative minBFs to the "folded normal alternative" bound (6). This appears to be true for Liebermeister's p-value and the mid p-value (see Figure 4), but the bound (6) does not seem to be reached for the more conservative standard p-values from Fisher's exact test. Accordingly, for $n \geq 300$, the estimated coefficients in model (23) are similar to the coefficients for bound (6) for the less conservative alternatives only (see Table S2 in the Supporting Information).

6.4 Results for one-sided simple alternatives

For most pairs of type of one-sided p-value and sample size considered, the values of R^2 were higher for model (22) than for model (23), especially for large sample sizes

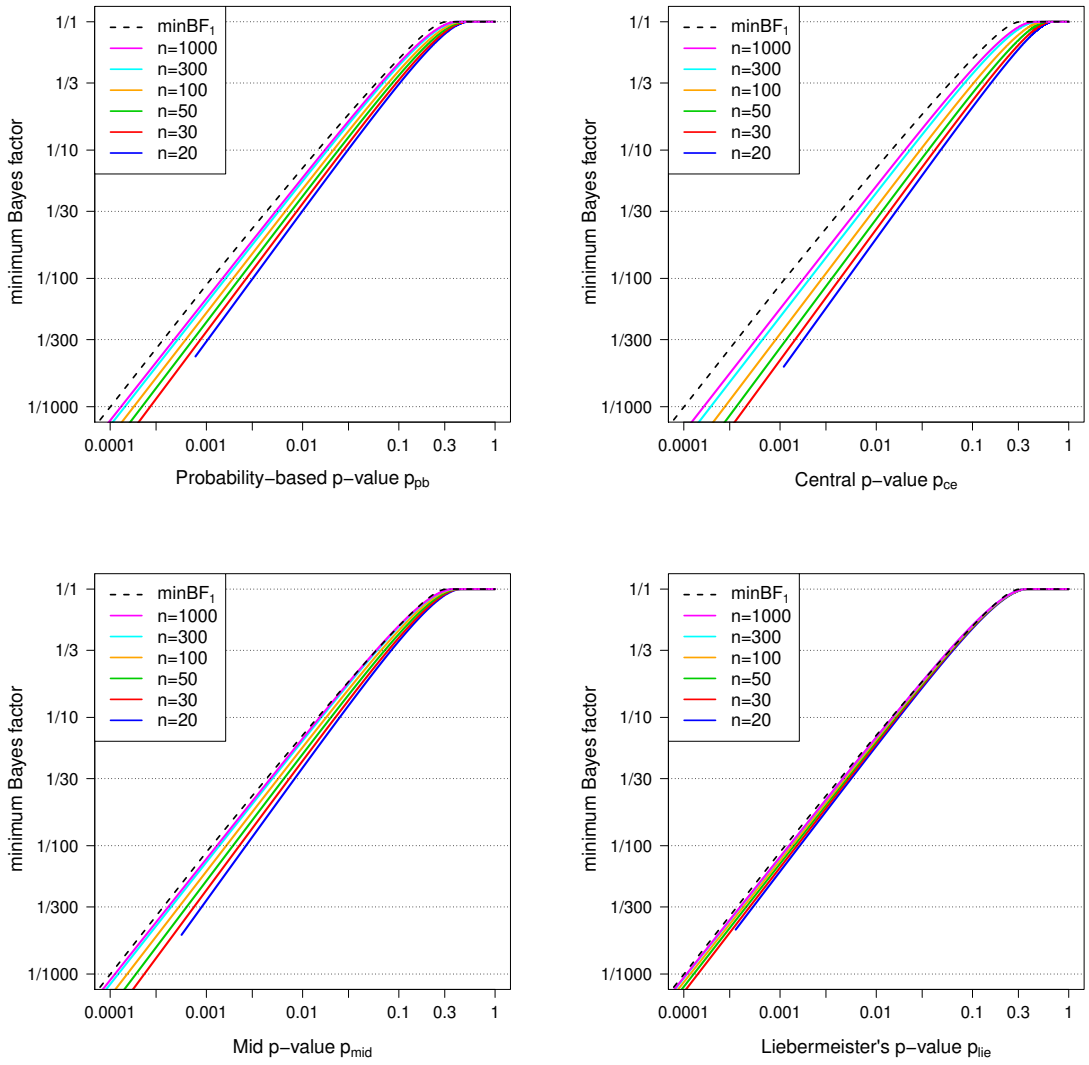

Figure 4: The colored curves show the estimated mean relationship (23) between two-sided p-values and two-sided simple alternative minBFs (20) for sample size $n = 20, 30, 50, 100, 300, 1000$, using the probability-based p-value (top left), the central p-value (top right), the mid p-value (bottom left) and Liebermeister's p-value (bottom right). For comparison, the large-sample "folded normal alternative" bound (6) is also displayed as a dashed line.

(see Tables S4 and S5 in the Supporting Information). Hence, we select model (22) for all analyses of one-sided simple alternative minBFs (21).

The estimated coefficients in model (22) and the corresponding fitted curves are

shown in the Supporting Information (Table S6 and Figure S4). It turns out that for all three types of one-sided p-values, the fitted minBFs are quite similar to the ones for the corresponding two-sided case shown in Figure 4. This is not surprising since in the one-sided case, both the p-value and the minBF are (approximately) half as large as in the two-sided case for all 2×2 tables with an odds ratio that is not close to 1. A similar correspondence between one- and two-sided simple alternatives has also been found for the two-sample t-test (Held & Ott, 2018, section 3.1).

6.5 Comparison with the minimum Bayes factor in the linear model

Interestingly, the fitted minBFs increase with increasing sample size for all types of non-asymptotic p-values considered (see Figures 3 and 4), so we observe the same qualitative relationship between sample size and minBFs as in the linear model (Held & Ott, 2016). For the class of local normal priors, we compare the fitted minBFs for 2×2 tables to the sample-size adjusted minBFs for the linear model with one degree of freedom ($d = 1$). In this special case, the F-test is equivalent to the standard two-sided two-sample t-test. The minBFs for the linear model turn out to be larger than the fitted minBFs for 2×2 tables based on any of the non-asymptotic p-values. While this is not surprising for the classical p-values from Fisher's exact test due to the conservativeness of the test, it is an interesting result for the mid p-value and Liebermeister's p-value, which are less conservative. A useful rule of thumb seems to be the following: If the sample size n of the 2×2 table is about four times as large as in the linear model, then the fitted minBFs based on Liebermeister's p-value are approximately equal to the sample-size adjusted minBFs for the linear model. The value $n/4$ for the effective sample size of 2×2 tables has also been suggested in Spiegelhalter et al. (2004, section 2.4.1), see Sabanés Bové & Held (2011, table 1) for a related derivation of the variance inflation factor 4 in logistic regression models.

Summary points

1. The maximal evidence of the p-values from Fisher's exact test is inversely related to sample size. This holds both under local normal alternatives as well as one- and two-sided simple alternatives.

2. The calibrations of p-values from Fisher's exact test exhibit a conservative bias even for large sample size. In contrast, the calibrations of the less conservative alternatives to these p-values - a mid p-value and the significance measure from Liebermeister's test - tend to the large-sample bound as the sample size becomes large.

7 Discussion

We have proposed sample-size adjusted minBFs for 2×2 contingency tables and related them to p-values from Fisher's exact test, as well as to less conservative alternative p-values. It turned out that on average (over all considered 2×2 tables of fixed sample size n), the maximal evidence of such non-asymptotic p-values against the point null hypothesis is inversely related to the sample size.

We would like to emphasize that the parametric models (22) and (23) relating p-values to minBFs are not intended to calibrate p-values for single 2×2 tables - in this setting, formula (18), (20) or (21) should be applied directly as they are more accurate. The purpose of these parametric models is to capture how the strength of evidence of p-values depends on the sample size. However, one limitation of our approach is that we cannot directly obtain sample-size adjusted minBFs for 2×2 tables with at least one entry equal to zero. In such cases, we recommend to first compute one of the available p-values and transform it using the calibration (22) or (23), respectively, with the corresponding coefficients (see Table 5 for the local alternative minBFs

and Table S2/S6 in the Supporting Information for the two-sided/one-sided simple alternative minBFs).

For 2×2 tables, the "$-ep\log(p)$" bound (3) is not "a best-case scenario for the strength of the evidence in favor of H_1 that can arise from a given p-value" (Bayarri et al., 2016). In fact, the sample-size adjusted local alternative minBFs (18) - which are tight lower bounds on the Bayes factor under normal priors - are on average smaller than the bound (3) for sample sizes $n \leq 50$ and all considered non-asymptotic p-values. For the 2×2 table with sample size 31 analyzed in Section 3.4, the sample-size adjusted minBFs (18) and (20) are also smaller than bound (3) for all types of p-values. Thus, we do not recommend to transform two-sided p-values from Fisher's exact test to the "$-ep\log(p)$" minBF (3), as this bound tends to be too large.

For the standard p-values from Fisher's exact test, we observed unfavorable large-sample behavior of the proposed calibrations for all three classes of prior distributions (see Figures 3, 4 and S4). Even for large sample size such as $n = 1000$, the calibrations based on these p-values still seem to suffer from a conservative bias, which is surprising. The large-sample behavior of the calibrations is best for Liebermeister's p-value, followed by the mid p-value. We thus propose to use one of these two p-values rather than the standard p-values from Fisher's exact test.

As we have seen, the minBFs obtained depend on the class of prior distributions over which the maximization is performed. Unfortunately, there is no "objective" choice for that class (Berger & Delampady, 1987). Although the local alternative and simple alternative minBFs differ, they do not do so by orders of magnitude. Some robustness of the minBFs with respect to different classes of prior distributions has also been observed in the multinomial (Delampady & Berger, 1990; Berger & Delampady, 1987) and the normal case (Berger & Sellke, 1987).

Our method to derive sample-size adjusted minBFs based on a class of generalized g-priors, which is an application of the Li & Clyde (2016) approach, can also be ap-

plied to other GLMs. For example, $m \times l$ contingency tables could also be handled. For Poisson regression with one binary covariate, a closed-form expression for the approximate minBF can also be obtained since the MLEs are available in closed form. Moreover, "non-asymptotic" (*i. e.* based on the exact distribution of the data) p-values also exist in this setting (Fay, 2010b). For other GLMs, however, the MLEs may need to be computed by numerical techniques such as the Newton-Raphson method or Fisher scoring.

In this paper, we have focused on tests of a point null hypothesis, also called tests for existence (Marsman & Wagenmakers, 2017). In tests for direction, the null and the alternative hypothesis are composite, *e. g.* H_0: $\beta < 0$ and H_1: $\beta > 0$. In such tests, the one-sided p-value is often equal or approximately equal to the posterior probability of H_0 if a non-informative prior is used (Casella & Berger, 1987), so that Bayesian calibrations of p-values are not necessary. For example, we have seen that the one-sided posterior probability (8) analyzed by Liebermeister corresponds to a one-sided p-value from Fisher's exact test for a slightly modified 2×2 table.

Acknowledgments

This work was supported by the Swiss National Science Foundation [project #159715]. We thank Isaac Gravestock for help with parallelizing `R`-code and two referees for helpful comments on an earlier version of this article.

References

Altham, P. M. E. (1969). Exact Bayesian analysis of a 2 x 2 contingency table, and Fisher's "exact" significance test. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **31**(2), 261–269.

Andrés, A. M. & Tejedor, I. H. (1997). On conditions for validity of the approximation to Fisher's exact test. *Biom. J.*, **8**, 935–954.

Barnard, G. A. (1989). On alleged gains in power from lower *P*-values. *Stat. Med.*, **8**, 1469–1477.

Bayarri, M. J., Benjamin, D. J., Berger, J. O. & Sellke, T. M. (2016). Rejection odds

and rejection ratios: A proposal for statistical practice in testing hypotheses. *J. Math. Psych.*, **72**, 90–103.

Bayarri, M. J. & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statist. Sci.*, **19**(1), 58–80.

Bayarri, M. J., Berger, J. O., Forte, A. & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.*, **40**(3), 1550–1577.

Berger, J. O. & Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.*, **2**(3), 317–335.

Berger, J. O. & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence (with discussion). *J. Amer. Statist. Assoc.*, **82**(397), 112–139.

Bland, M. (2015). *An Introduction to Medical Statistics*. 4th ed. Oxford University Press.

Breslow (1981). Odds ratio estimators when the data are sparse. *Biometrika*, **68**(1), 73–84.

Breslow, N. & Powers, W. (1978). Are there two logistic regressions for retrospective studies?. *Biometrics*, **34**, 100–105.

Casella, G. & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.*, **82**(397), 106–111.

D'Agostino, R. B., Chase, W. & Belanger, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations. *Amer. Statist.*, **42**(3), 198–202.

Davidson, R. R. & Lever, W. E. (1970). The limiting distribution of the likelihood ratio statistic under a class of local alternatives. *Sankhya A*, **32**(2), 209–224.

Davison, A. C. (2003). *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Delampady, M. & Berger, J. O. (1990). Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit. *Ann. Statist.*, **18**(3), 1295–1316.

Di Sebastiano, P., Fink, T., Di Mola, F. F., Weihe, E., Innocenti, P., Friess, H. & Büchler, M. W. (1999). Neuroimmune appendicitis. *Lancet*, **354**, 461–466.

Dunne, A., Pawitan, Y. & Doody, L. (1996). Two-sided P-values from discrete asymmetric distributions based on uniformly most powerful unbiased tests. *J. Roy. Statist. Soc. Ser. D (The Statistician)*, **45**(4), 397–405.

Edwards, W., Lindman, H. & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.*, **70**(3), 193–242.

Engle, R. F. (1984). *Wald, likelihood ratio, and Lagrange multiplier tests in econometrics*. Vol. 2 of *Handbook of Econometrics*. North-Holland. Chapter 13, pp. 775–826.

Fay, M. P. (2010a). Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostatistics*, **11**(2), 373–374.

Fay, M. P. (2010b). Two-sided exact tests and matching confidence intervals for discrete data. *The R Journal*, **2**(1), 53–58.

Fisher, R. A. (1941). *Statistical Methods for Research Workers*. Oliver & Boyd.

Ghosh, J., Purkayastha, S. & Samanta, T. (2005). Role of P-values and other measures of evidence in Bayesian analysis. *In* D. K. Dey & C. R. Rao (eds), *Bayesian Thinking: Modeling and Computation*. Vol. 25 of *Handbook of Statistics*. Elsevier. Chapter 5, pp. 151–170.

Goodman, S. N. (1992). A comment on replication, *P*-values and evidence. *Stat. Med.*, **11**(7), 875–879.

Hansen, M. H. & Yu, B. (2003). Minimum description length model selection criteria for generalized linear models. *Lecture Notes-Monograph Series*, **40**, 145–163. Statistics and Science: A Festschrift for Terry Speed.

Held, L. & Ott, M. (2016). How the maximal evidence of *P*-values against point null hypotheses depends on sample size. *Amer. Statist.*, **70**(4), 335–341.

Held, L. & Ott, M. (2018). On *p*-values and Bayes factors. *Annu. Rev. Stat. Appl.*, **5**, 393–419.

Held, L., Sabanés Bové, D. & Gravestock, I. (2015). Approximate Bayesian model selection with the deviance statistic. *Statist. Sci.*, **30**(2), 242–257.

Hirji, K. F., Tan, S.-J. & Elashoff, R. M. (1991). A quasi-exact test for comparing two binomial proportions. *Stat. Med.*, **10**, 1137–1153.

Howard, J. V. (1998). The 2 x 2 table: a discussion from a Bayesian viewpoint. *Stat. Sci.*, **13**(4), 351–367.

Hwang, J. T. G. & Yang, M.-C. (2001). An optimality theory for mid *p*-values in 2×2 contingency tables. *Statist. Sinica*, **11**, 807–826.

Johnson, V. E. (2005). Bayes factors based on test statistics. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**(5), 689–701.

Johnson, V. E. (2008). Properties of Bayes factors based on test statistics. *Scand. J. Stat.*, **35**, 354–368.

Johnson, V. E. (2013). Revised standards for statistical evidence.. *Proc. Natl. Acad. Sci. USA*, **110**(48), 19313–19317.

Kass, R. & Vaidyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **54**(1), 129–144.

Kateri, M. (2014). *Contingency Table Analysis - Methods and Implementation using R*. Statistics for Industry and Technology. Birkhäuser.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *J. Amer. Statist. Assoc.*, **56**(294), 223–234.

Li, Y. & Clyde, M. A. (2016). Mixtures of g-priors in generalized linear models. arXiv:1503.06913v2 [stat.ME].

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.*, **103**(481), 410–423.

Liebermeister, C. (1877). Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. Sammlung klinischer Vorträge. *Innere Medizin*, **110**(31–64), 935–962.

Lloyd, C. J. (1988). Doubling the one-sided P-value in testing independence in 2×2 tables against a two-sided alternative. *Stat. Med.*, **7**, 1297–1306.

Ly, A. (2017). *Bayes Factors for Research Workers*. PhD thesis. University of Amsterdam.

Marsman, M. & Wagenmakers, E.-J. (2017). Three insights from a Bayesian interpretation of the one-sided P value. *Educ. Psychol. Meas.*, **77**(3), 529–539.

Meulepas, E. (1998). A two-tailed P-value for Fisher's exact test. *Biom. J.*, **40**(1), 3–10.

Nurminen, M. & Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scand. J. Stat.*, **14**, 67–77.

Overall, J. E. (1980). Continuity correction for Fisher's exact probability test. *Journal of Educational Statistics*, **5**(2), 177–190.

Rothman, K. J. & Greenland, S. (1998). *Modern Epidemiology*. 2nd ed. Lippincott-Raven.

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *Amer. Statist.*, **40**(4), 313–315.

Sabanés Bové, D. & Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Anal.*, **6**(3), 387–410.

Sellke, T., Bayarri, M. J. & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. Statist.*, **55**(1), 62–71.

Seneta, E. (1994). Carl Liebermeister's hypergeometric tails. *Historia Math.*, **21**, 453–462.

Seneta, E. & Phipps, M. C. (2001). On the comparison of two observed frequencies. *Biom. J.*, **43**(1), 23–43.

Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley.

Vovk, V. G. (1993). A logic of probability, with application to the foundations of statistics (with discussion and a reply by the author). *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **55**(2), 317–351.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, **14**(5), 779–804.

Wang, X. & George, E. I. (2007). Adaptive Bayesian criteria in variable selection for generalized linear models. *Statist. Sinica*, **17**(2), 667–690.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *In* P. K. Goel & A. Zellner (eds), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Vol. 6 of *Studies in Bayesian Econometrics and Statistics*. North-Holland. Amsterdam. Chapter 5, pp. 233–243.

Supporting information

Additional information can be found in the supplementary PDF file provided.